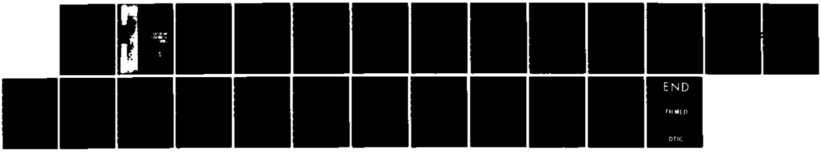
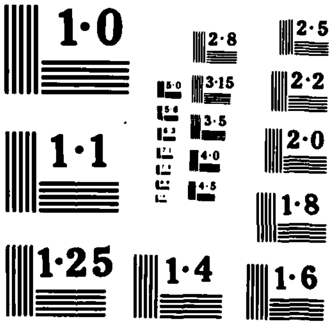
AD-A152 336 THE ROLE OF DUALITY IN OPTIMIZATION PROBLEMS INVOLVING ENTROPY FUNCTIONAL..(U) TEXAS UNIV AT AUSTIN CENTER FOR CYBERNETIC STUDIES A BEN-TAL ET AL. SEP 84 CCS-497 1/1

UNCLASSIFIED N00014-82-K-0295 F/G 5/2 NL




1·0
2·8
2·5
5·0
5·6
3·15
2·2
3·5
1·1
2·0
4·0
4·5
1·8
1·25
1·4
1·6

2

Research Report CCS 497

THE ROLE OF DUALITY IN OPTIMIZATION PROBLEMS INVOLVING ENTROPY FUNCTIONALS, WITH APPLICATIONS TO INFORMATION THEORY

by

A. Ben-Tal*
M. Teboulle*
A. Charnes

CENTER FOR CYBERNETIC STUDIES
The University of Texas
Austin, Texas 78712

6

Research Report CCS 497

THE ROLE OF DUALITY IN OPTIMIZATION PROBLEMS INVOLVING ENTROPY FUNCTIONALS, WITH APPLICATIONS TO INFORMATION THEORY

by

A. Ben-Tal*
M. Teboulle*
A. Charnes

September 1984

*The Technion-Israel Institute of Technology

This paper was partly supported by ONR Contract N00014-81-C-0236 and N00014-82-K-0295 with the Center for Cybernetic Studies, The University of Texas at Austin.
Research partly supported by NSF Grant No. ECS-8214081, Technion VPR-Fund, Andre Meyer Fund for Operation Research and Fund for the Promotion of Research at the Technion.
Reproduction in whole or in part is permitted for any purpose of the United States Government.

CENTER FOR CYBERNETIC STUDIES

A. Charnes, Director
College of Business Administration 5.202
The University of Texas at Austin
Austin, Texas 78712-1177
(512) 471-1821

ABSTRACT

We consider infinite dimensional optimization problems involving entropy-type functionals in the objective function as well as in the constraints. A duality theory is developed for such problems and applied to the reliability rate function problem in Information Theory.

Key Words: Optimization in Infinite Dimensional Spaces; Duality in Convex Optimization; Entropy; Divergence; Information Theory; Channel Capacity; Reliability Rate Function; Error Exponent Function.

1. INTRODUCTION

Extremum problems involving entropy-type functionals appear in a diversity of applications. To mention just few: statistical estimation and hypothesis testing (Kullback-Leibler (Ref. 1), Kullback (Ref. 2), Akaike (Ref. 3)), traffic engineering (Charnes et al. (Ref. 4)), marketing (Charnes et al (Ref. 5)), accounting (Charnes and Cooper (Refs. 6,7)), information theory (Shannon (Ref. 8)).

In the majority of these applications, the extremum problems involved are studied only for the case of finite distributions. Extensions to arbitrary distributions were derived recently by Ben-Tal and Charnes (Ref. 9). The extremum problem is set up as an infinite dimensional convex program with linear equality constraints, namely:

$$\text{(A)} \quad \inf_{f \in D} \left\{ \int_T f(t) \log \frac{f(t)}{g(t)} dt : \int_T f(t) a_i(t) dt = \theta_i \qquad i = 1, \dots, m \right\},$$

where D is the convex subset of density functions with support T, and $g(\cdot)$ is a given density in D.

It is shown in Ref. 9 that the dual problem is the <u>unconstrained finite dimensional</u> concave program:

$$\text{(B)} \quad \sup_{y \in \mathbb{R}^m} \left\{ y^t \theta - \log \int_T g(t) e^{\sum_{i=1}^{m} y_i a_i(t)} dt \right\}$$

The dual pair (A)-(B) has a very interesting statistical interpretation: let $\{\theta_i\}_{i=1}^m$ be parameters of the distribution, estimated in terms of a sample $x = (x_1, \dots, x_n)$ by

$$\hat{\theta}_i(x) = \hat{\theta}_i(x_1, \dots, x_n) = \frac{1}{n} (a_i(x_1) + \dots + a_i(x_n))$$

and let these estimates replace θ_i in the constraints of (A).

Consider now the problems of finding the maximum likelihood estimator $\pi^*(x)$ of the parameter vector $\pi = (\pi_1, \ldots, \pi_m)^t$ in the exponential family generated by the (fixed) density $g(t)$, i.e.:

$$f(t|\pi) = g(t)c(\pi)e^{\sum_{i=1}^{m} \pi_i a_i(t)}$$

where $c(\pi)$ is a normalizing constant, i.e.:

$$c(\pi)^{-1} = \int_T g(t)e^{\sum_{i=1}^{m} \pi_i a_i(t)} dt.$$

The likelihood function is

$$\prod_{j=1}^{n} f(x_j|\pi) = \{\prod_{j=1}^{n} g(x_j)\} \cdot c(\pi)^n e^{\sum_j \sum_i \pi_i a_i(x_j)}$$

hence $\frac{1}{n}$ log(likelihood) = const. + $\log\left[c(\pi)e^{\sum \pi_i \hat{\theta}_i(x)}\right]$ therefore the maximum likelihood estimator $\pi^*(x)$ is obtained by solving:

$$\max_{\pi \in \mathbb{R}^m} \{\sum \pi_i \hat{\theta}_i(x) - \log c^{-1}(\pi)\} =$$

$$\max_{\pi \in \mathbb{R}^m} \{\sum_{i=1}^{m} \pi_i \hat{\theta}_i(x) - \log \int_T g(t)e^{\sum_{i=1}^{m} \pi_i a_i(t)}\}$$

The latter is precisely the dual problem (B). Thus, for the exponential family, Statistical Information theory and the Maximum Likelihood approach are dual principles.

Many problems in information theory, however, cannot be stated just with linear constraints as in problem (A), they contain also (nonlinear) entropy type inequality constraints. It is the purpose of this paper to derive duality results for such problems and to demonstrate their power and elegance in treating such problems.

As a motivation we begin by describing the channel capacity problem of Information Theory. Consider a communication channel described by an input alphabet $A = \{1,\dots,n\}$, an output alphabet $B = \{1,\dots,m\}$ and by a probability transition matrix $Q = \{Q(k|j)\}$, where $Q(k|j)$ is the probability of receiving the output letter $k \in B$ when input letter $j \in A$ was transmitted.

The capacity of the channel is defined as:

$$C = \max_{p \in P^n} (I(p,Q) \overset{\Delta}{=} \max_{p \in P^n} \sum_{k=1}^{m} \sum_{j=1}^{n} p_j Q(k|j) \log \frac{Q(k|j)}{\sum_{\ell=1}^{n} p_\ell Q(k|\ell)} \qquad (1)$$

where

$$\mathbb{P}^n \overset{\Delta}{=} \{p \in \mathbb{R}^n : p_j \geq 0 \ \forall j; \ \sum_{j=1}^{n} p_j = 1\} \qquad (2)$$

is the set of all probability distributions on the channel input, and $I(p,Q)$ is known as the average mutual information between the channel input and channel output. Channel capacity is the basic concept of Shannon's mathematical theory of communication (later called Information theory). For more details on the notion of capacity and its significance, the reader is referred to Shannon (Ref. 8), Gallager (Ref. 10), Jelinek (Ref. 11).

Roughly speaking, the basic theorem of information theory, the so-called "noisy channel coding theorem", states that if the channel has capacity C, it is possible to transmit over this channel messages, of sufficiently large length, at rate $R < C$ and still be able to decode them with a probability of error as small as desired. Upper bound on the probability of error is given in terms of an exponential decreasing function of the so-called reliability rate function $E(R)$. In the classical proof of the coding theorem, the function E(R) is derived via a sequence of mathematical manipulations, see e.g., Gallager (Ref. 12) and Csiszar (Ref. 13). Blahut (Ref. 14), has enlightened many basic problems of coding theory by defining E(R) as a saddle function problem, involving the Kullback-Leibler relative entropy functional, namely, for a given channel matrix $P(k|j)$:

$$E(R) = \max_{p\in P^n} \min_{Q\in\mathcal{Q}(R)} \sum_{k=1}^{m} \sum_{j=1}^{n} p_j Q(k|j) \log \frac{Q(k|j)}{P(k|j)} \tag{3}$$

where

$$\mathcal{Q}(R) = \{Q\colon\ I(p,Q) \leq R\}, \quad R \text{ a positive scalar.}$$

Starting from this definition, Blahut (Ref. 14) proved, that E(R) can be expressed by the conventional parametric form originally proposed by Gallager (Ref. 12) namely,

$$E(R) = \max_{\delta\geq 0} \max_{p\in \mathbb{P}^n} \{-\delta R - \log \sum_{k=1}^{m} \{ \sum_{j=1}^{n} p_j P(k|j)^{\frac{1}{1+\delta}} \}^{1+\delta}\} \tag{4}$$

A new proof of this result is given here in Section 3, via the duality theory developed in Section 2. The duality framework can be applied to a variety of other extremum problems of information theory, (see e.g., Blahut (Ref. 14), Table I, p. 417).

In particular, more than one entropy-type constraint can be easily dealt with, and the general (not necessarily discrete) distribution case can be considered.

2. DUALITY THEORY FOR LINEAR AND ENTROPY CONSTRAINED PROGRAMS

Let dt be a σ-finite additive measure defined on a σ-field of the subsets of a measurable space T, and let $L^1 \overset{\Delta}{=} L^1(T,dt)$ be the usual Lebesgue space of measurable real valued functions x on T so that

$$\|x\| \overset{\Delta}{=} \int_T |x(t)|dt < \infty .$$

Let $\mathbb{D} = \{x \in L^1 : x(t) \geq 0(a.e), \int_T x(t)dt = 1\}$ be the convex subset of L^1 which is the set of all probability densities $x(\cdot)$ on T.

Consider the infinite dimensional optimization problem:

$$\text{(P)} \quad \inf \int_T x(t) \log \frac{x(t)}{c_o(t)} dt$$

subject to

$$\int_T a_i(t)x(t)dt \geq b_i \qquad i \in I \overset{\Delta}{=} \{1,\ldots,m\} \tag{5}$$

$$\int_T x(t) \log \frac{x(t)}{c_k(t)} \leq e_k \qquad k \in K \overset{\Delta}{=} \{1,\ldots,p\} \tag{6}$$

$$x(t) \in \mathbb{D} \subset L^1$$

where $c_k : T \to \mathbb{R}$, $k \in \{0\} \cup K$ are given summable positive functions; $a_i : T \to \mathbb{R}$ are given continuous functions; and $\{b_i\}_{i\in I}$, $\{e_k\}_{k\in K}$ are given real numbers.

Here and henceforth, $0 \log 0 = \lim_{t\to 0^+} t \log t = 0$. A dual representation of problem (P) will be derived via Lagrangian duality. Recall that for a convex optimization problem:

(A) $\quad \inf \{f(x) : g(x) \leq 0 \quad x \in C \subset X\}$

where $f: C \to \mathbb{R}$, $g: C \to \mathbb{R}^m$ are convex functions defined on a convex subset C of a linear space X, the Lagrangian for problem (A) is defined as $L: C \times \mathbb{R}^m_+ \to \mathbb{R}$ given by:

$$L(x,y) = f(x) + y^t g(x).$$

The dual objective function is

$$h(y) = \inf_{x \in C} L(x,y)$$

and then the dual problem (B) associated with (A) is defined as:

$$\text{(B)} \quad \sup_{y \geq 0} h(y).$$

The main result concerning the dual pair (A) and (B) is the existence of a saddle point (x^*,y^*) for (A) or equivalently, the validity of a strong duality result:

$$\inf(A) = \max(B)^{(*)}.$$

Under the familiar Slater regularity condition:

$$\text{(S)} \quad \exists x \in C: g(x) < 0$$

the strong duality relation is guaranteed. More precisely we have: (see e.g., Rockafellar (Refs. 15,16), Laurent (Ref. 17), Ponstein (Ref. 18))

Theorem 2.1 Assume that $\inf(A) < \infty$ and that the regularity assumption (S) holds then

$$\inf(A) = \max(B).$$

□

Remark 2.1 The regularity condition (S) is, in fact, related to the notion of stably set problem. More details are available

(*) We follow the convention of writing "min" ("max") if the infinum (supremum) is attained.

in Rockafellar (Ref. 15) and Laurent (Ref. 17) (especially Theorem 7.6.1, p. 403).

Remark 2.2 A result of the type of Theorem 1 has typically a symmetric version, i.e., if (B) is assumed stably set then min(A) = sup(B), (see Rockafellar (Ref. 15), Theorem 4, p. 179).

We now return to the primal "entropy problem" (P). The derivation of its dual objective function is based on the following simple result.

Lemma 2.1 Let $s(t)$ be a given positive summable function:

$$\int_T s(t)dt = S < \infty \quad .$$

Then

$$\min_{x \in \mathbb{D}} \int_T x(t) \log \frac{x(t)}{s(t)}\, dt = - \log S$$

where the optimal probability density $x^*(t) = \frac{s(t)}{S}$ [a.e].

Proof: Define $h(t) = \frac{s(t)}{S}$, then $h(t) \in \mathbb{D}$, hence by Theorem 3.1, p. 14, Kullback (Ref. 2), we have:

$$\inf_{x \in \mathbb{D}} \int_T x(t) \log \frac{x(t)}{h(t)}\, dt = 0$$

where the infimum is attained for $x^*(t) = h(t) = \frac{s(t)}{S}$.

Then using the identity:

$$\int_T f(t) \log \frac{f(t)}{s(t)}\, dt = \int_T f(t) \log \frac{f(t)}{h(t)} - \log \int_T s(t) dt$$

the result follows. □

The Lagrangian for problem (P) is $L: \mathbb{D} \times \mathbb{R}^m_+ \times \mathbb{R}^p_+ \to \mathbb{R}$

$$L(x,y,\lambda) = b^t y - e^t \lambda + \int_T \{\log \frac{x(t)}{c_o(t)} - \sum_{i \in I} y_i a_i(t) + \sum_{k \in K} \lambda_k \log \frac{x(t)}{c_k(t)} \} x(t) dt \tag{7}$$

and thus the dual problem (D) associated with (P) is defined as:

$$\sup\{\inf_{x \in D} L(x,y,\lambda) : y \in \mathbb{R}^m_+, \lambda \in \mathbb{R}^p_+ \} .$$

The next result shows that the dual problem (D) can be expressed simply as a finite dimensional concave program involving only nonnegative constraints.

Theorem 2.2 The dual problem of (P) is given by:

$$(D) \quad \sup_{\substack{y \in \mathbb{R}^m_+ \\ \lambda \in \mathbb{R}^p_+}} \{y^t b - \lambda^t e - \rho \log \int_T c_o(t) e^{\frac{1}{\rho}\{\lambda^t B(t) + y^t A(t)\}} dt\}$$

where:

$$\rho = 1 + \sum_{k=1}^{p} \lambda_k , \quad \text{and}$$

$$A(t) = (a_1(t),\ldots,a_m(t))^t ; \quad B(t) = (B_1(t),\ldots,B_p(t))^t$$

with

$$B_k(t) = \log \frac{c_k(t)}{c_o(t)} \qquad \forall k \in K = \{1,\ldots,p\}.$$

Proof: The Lagrangian defined in (7) can be written after some algebraic manipulations as:

$$L(x,y,\lambda) = -y^t b - \lambda^t e + \int_T x(t) \log \frac{x(t)^{(1+\sum_{k=1}^{p} \lambda_k)}}{(\prod_{k=0}^{p} c_k(t)^{\lambda_k}) e^{y^t A(t)}} dt .$$

Then, defining $\rho = 1 + \sum_{k=1}^{p} \lambda_k$ and $B_k(t) = \log \frac{c_k(t)}{c_o(t)}$, a little algebra shows that the dual objective function can be expressed as:

$$h(y,\lambda) = y^t b - \lambda^t e + \rho \inf_{x \in D} \int_T x(t) \log \frac{x(t)}{c_o(t) e^{\frac{1}{\rho}\{\lambda^t B(t) + y^t A(t)\}}} dt .$$

Now, applying Lemma 2.1 with $s(t) = c_o(t) e^{\frac{1}{\rho}\{\lambda B(t) + y^t A(t)\}}$ we get the desired result.

□

Duality results for the pair of problems (P)-(D) will now follow by setting problem (P) as a convex program of the type (A) and then applying Theorem 2.1.

Theorem 2.3

(a) If (P) is feasible then inf(P) is attained and

min(P) = sup(D).

Moreover, if there exists $x \in D$ satisfying the constraints (5), (6) strictly, then sup(D) is attained and

min(P) = max(D).

(b) If $x^* \in D$ solves (P) and $y^* \in \mathbb{R}^m_+$, $\lambda^* \in \mathbb{R}^p_+$ solves (D) then:

$$x^*(t) = \frac{c_o(t) e^{\frac{1}{\rho}\{\lambda^{*t} B(t) + y^{*t} A(t)\}}}{\int_T c_o(t) e^{\frac{1}{\rho}\{\lambda^{*t} B(t) + y^{*t} A(t)\}} dt} \quad [a.e]$$

Proof: In order to apply Theorem 2.1, we need to set problem (P) in the format of the convex program (A). Thus, consider the linear operator $A: L^1 \to \mathbb{R}^m$ given by:

$$x \to \begin{pmatrix} \int_T a_1(t)x(t)dt \\ \vdots \\ \int_T a_m(t)x(t)dt \end{pmatrix}$$

and for $k \in \{0\} \cup K$, define the integral functionals

$$I_k(x) = \begin{cases} \int_T x(t) \log \frac{x(t)}{c_k(t)} dt & \text{if } x \in \mathbb{D} \\ \infty & \text{otherwise.} \end{cases}$$

Then problem (P) can be written as a convex optimization problem.

$$\text{(P)} \quad \inf\{I_o(x) : Ax \geq b, \; I_k(x) \leq e_k \;\; k \in K, \; x \in \mathbb{D}\}.$$

Note that (P) corresponds to (A) with

$$X := L^1, \quad C := \mathbb{D}, \quad f(x) := I_o(x) \quad \text{and} \quad g(x) := \begin{pmatrix} b-Ax \\ I_1(x)-e_1 \\ \vdots \\ I_p(x)-e_p \end{pmatrix}$$

and then the results follow from Theorem 2.1. In fact, since the dual (D), given in Theorem 2.2, has only nonnegative constraints ($y \geq 0$, $\lambda \geq 0$), it satisfies the strongest constraint, implying by Remark 2.2 lack of duality gap and attainment of the primal infimum. Thus the first part of conclusion (a) follows. The second part follows directly from Theorem 2.1 itself. Moreover, part (a) implies the existence of a saddle point $(x^*(t), y^*, \lambda^*) \in \mathbb{D} \times \mathbb{R}^m_+ \times \mathbb{R}^p_+$, so

$$\min_{x \in \mathbb{D}} L(x, y^*, \lambda^*) = L(x^*, y^*, \lambda^*)$$

and the expression for x^* given in (b) follows from the last part of Lemma 2.1.

□

3. AN APPLICATION IN INFORMATION THEORY

In this section we apply the duality relation for problem (P) to treat in a unified simple way the reliability rate function problem described in the introduction. While our results developed in Section 2 are applicable to the case of general probability distributions, we restrict ourselves here to the case of finite discrete probability distributions, since they include most of the interesting problems appearing in information theory. We begin with some further notations and definitions, following closely the terminology of (Ref. 14).

The relative entropy or discrimination between two discrete (finite) distributions p, q playing a fundamental role in statistical information theory is a function $J: \mathbb{P}^n \times \mathbb{P}^n \to \mathbb{R}$ defined by:

$$J(p,q) = \sum_{k=1}^{n} p_k \log \frac{p_k}{q_k} \tag{8}$$

It is well known that J is convex in each of its arguments, nonnegative, and equal to zero if and only if $p_k = q_k$, $\forall k$, (see e.g., Ref.2).

Similarly, one defines the average discrimination by:

$$J(Q,P) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_j Q(k\ j) \log \frac{Q(k|j)}{P(k|j)} \tag{9}$$

where p, Q, P are as defined in the Introduction.

In the rest of this paper we simplify the notations: probability transition matrix like $P(k|j)$ are denoted P_{kj} and summation indices are dropped.

An error exponent function is defined in Blahut [1974] as the following (single) entropy constrained program:

(E) $e(r) = \min\{J(q,q_2) : q \in \mathbb{P}(r)\}$

where

$$\mathbb{P}(r) = \{q \in \mathbb{P}^n : J(q,q_1) \leq r\} \quad .$$

r is a given positive scalar and q_1, q_2 are given distributions in $\mathbb{P}^n$. Problem (E) just defined is a special case of problem (P), described in Section 2 with: $I = \emptyset$ (i.e., no linear constraints), $K = \{1\}$ and with $c_o(t)$, $c_1(t)$ corresponding here to the discrete finite distributions q_2, q_1 respectively. Moreover, since problem (E) consists of minimizing continuous function over the compact set $\mathbb{P}(r)$, the minimum is attained; we know also from Theorem 2.2 that the dual problem (H) corresponding to (E) involves only nonnegative constraints, hence satisfying the strongest constraint qualifications; we get according to Theorem 2.2 and Theorem 2.3, by setting

$$\rho = 1 + \lambda_1 = 1 + \delta \qquad \text{and} \quad e_1 = r,$$

<u>Theorem 3.1</u> A dual representation of (E) is the program

$$\text{(H)} \qquad e(r) = \max_{\delta \geq 0} \left\{ -\delta r - \log\left(\sum q_{1k}^{\frac{\delta}{1+\delta}} \, q_{2k}^{\frac{\delta}{1+\delta}} \right)^{1+\delta} \right\} \quad .$$

Moreover, if $q^* \in \mathbb{P}^n$ solves (E) and $\delta^* \geq 0$ solves (H) then

$$q_k^* = \frac{q_{1k}^{\frac{\delta}{1+\delta}} \, q_{2k}^{\frac{1}{1+\delta}}}{\sum_k q_{1k}^{\frac{\delta}{1+\delta}} \, q_{2k}^{\frac{1}{1+\delta}}}$$

□

We recover here, a result obtained in ((Ref.14), Theorem 7).

We now derive the dual representation of E(R) by reference to the error exponent function e(r).

Recalling the definition of the reliability rate function given in the Introduction (see eq. (3)) and using our notations we have:

$$E(R) = \max_{p \in \mathbb{P}^n} \min_{Q \in \mathcal{Q}(R)} J(Q,P) \tag{10}$$

where

$$\mathcal{Q}(R) = \{Q: I(p,Q) \leq R\}.$$

A useful identity for the average mutual information is

$$I(p,Q) = \min_{q \in \mathbb{P}^n} \bar{J}(Q,q) \tag{11}$$

where

$$\bar{J}(Q,q) := \sum_{k=1}^{m} \sum_{j=1}^{n} p_j Q(k|j) \log \frac{Q(k|j)}{q_k} \quad ,$$

this can be verified by observing that the minimum is achieved for $q_k^* = \sum_j p_j Q_{kj}$.

Using (11), problem (10) can be reformulated as:

$$E(R) = \max_{p \in \mathbb{P}^n} \min_{Q \in \mathcal{Q}(R)} \{J(Q,P): \min_{q \in \mathbb{P}^n} \bar{J}(Q,q) \leq R\} . \tag{12}$$

Now it is an easy exercise to show that any optimization problem of the form $\min_x \{f(x): \min_y g(x,y) \leq r\}$ is equivalent to $\min_{x,y} \{f(x): g(x,y) \leq r\}$ hence (12) becomes

$$E(R) = \max_p \min_q \min_Q \{J(Q,P): \bar{J}(Q,q) \leq R\} \quad . \tag{13}$$

The inner minimum in (13) is of the form of $e(r)$ in problem (E), and is appropriately denoted by $e(R,q)$. Then by Theorem 3.1, a dual representation of it is easily shown to be:

$$e(R,q) = \max_{\delta \geq 0} \{-\delta R - \log \left\{ \sum_k \sum_j p_j P_{kj}^{\frac{1}{i+\delta}} q_k^{\frac{\delta}{1+\delta}} \right\}^{1+\delta} \tag{14}$$

Substituting the latter representation in (13), we get

$$E(R) = \max_{p} \min_{q} \max_{\delta\geq 0} \{g(q,\delta) - \delta R\} \tag{15}$$

where

$$g(q,\delta) := -\log \left\{ \sum_{k} \sum_{j} p_j P_{kj}^{\frac{1}{1+\delta}} \, q^{\frac{\delta}{1+\delta}} \right\}^{1+\delta} \tag{16}$$

We shall prove that the "min-max" appearing in (15) can be reversed. Before, we need an auxiliary result.

Lemma 3.1 The function $g(q,\delta)$ defined in (16) is

(a) concave in δ for any $q \in \mathbb{P}^n$

(b) convex in q for any $\delta \geq 0$.

Proof: (a) It is well known that the Lagrangian dual function is always concave in the dual variables, hence (a) follows.

(b) Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex decreasing function, and let $g\colon \mathbb{R}^n \to \mathbb{R}$ be a concave function, then it is easy to verify that $h(x) = f(g(x))$ is convex.

Take $f(t) = -\log t$ (convex decreasing), $g(q) = \sum_k a_k q_k^{\frac{\delta}{1+\delta}}$ with $a_k := \sum_j p_j P_{kj}^{\frac{\delta}{1+\delta}} > 0$ (concave for $\delta \geq 0$) then clearly $g(q,\delta) = (1+\delta) f(g(q))$ and (b) is proved.

□

The min-max theorem related to (15) now follows.

Theorem 3.1 Let $K(q,\delta) = g(q,\delta) - \delta R$

$$\min_{q} \max_{\delta\geq 0} K(q,\delta) = \max_{\delta\geq 0} \min_{q} K(q,\delta) \tag{17}$$

Proof: By Lemma 3.1, $K(q,\delta)$ is a convex-concave saddle function for every $q \in \mathbb{P}^n$ and every $\delta \geq 0$. By a result of Rockafeller (Ref. 19), a sufficient condition for the validity of (17) for a general convex-concave saddle function is:

$$\begin{cases} \nexists \delta_o \geq 0 \quad \text{such that} \\ \delta_o \dfrac{dK}{d\delta}(q,\delta) \geq 0 \quad (q \in \mathbf{P}^n,\ \delta > 0). \end{cases}$$

This is certainly satisfied if:

$$\exists\, q,\ \exists\ \delta > 0 \quad \text{such that} \quad \frac{dK}{d\delta}(q,\delta) < 0,$$

i.e.,

$$\exists q,\ \exists \delta > 0:\ g'(q,\delta) = \frac{d}{d\delta}\, g(q,\delta) < R \quad . \tag{18}$$

Since $R > 0$, it suffices to prove that:

$$\inf_{\delta \geq 0} g'(q,\delta) \leq 0 \ . \tag{19}$$

But $g'(q,\delta)$ is a derivative of a concave function and thus is decreasing, hence

$$\inf_{\delta \geq 0} g'(q,\delta) = \lim_{\delta \to \infty} g'(q,\delta) \quad . \tag{20}$$

Moreover, the gradient inequality for the concave function $g(q,\cdot)$ implies:

$$0 = g(q,0) \leq g(q,\delta) - \delta g'(q,\delta)$$

hence:

$$g'(q,\delta) \leq \frac{g(q,\delta)}{\delta} \quad .$$

Thus, to prove (19) it suffices to show that

$$\lim_{\delta \to \infty} \frac{g(q,\delta)}{\delta} \leq 0 \quad .$$

Indeed, straightforward computation shows that $\lim\limits_{\delta \to \infty} \dfrac{g(q,\delta)}{\delta} = 0$.

□

The last theorem permits us to write E(R) (see, eq. (15)) as:

$$E(R) = \max_{p} \max_{s \geq 0} \min_{q} K(q,\delta) \quad .$$

However, the next result, will show that the inner minimum can be computed, and thus E(R) can be expressed simply as a double maximum problem.

Lemma 3.2

$$\max_{x \in X} \log \left(\sum x_i^{\frac{\alpha}{1+\alpha}} y_i\right)^{1+\alpha} = \log \sum y_i^{1+\alpha} \qquad (\alpha > 0)$$

where

$$X = \{x \in \mathbb{R}^n : \ x_k \geq 0 \ \sum_{k=1}^{n} x_k = 1\} \ .$$

Proof: From Holder inequality we get

$$\left(\sum x_k^{\frac{\alpha}{1+\alpha}} y_k\right)^{1+\alpha} \leq \left(\sum x_k\right)^{\alpha}\left(\sum y_k^{1+\alpha}\right) \ .$$

Taking log of both expressions and using the fact that $\sum x_k = 1$, we get:

$$\sup_{x \in X} \log\left(\sum x_k^{\frac{\alpha}{1+\alpha}} y_k\right)^{1+\alpha} \leq \log \sum y_k^{1+\alpha}$$

and the sup is attained for $x_k^* = \dfrac{y_k^{1+\alpha}}{\sum y_k^{1+\alpha}}$.

□

Now, since

$$\min_{q} K(q,\delta) = -\delta R - \max_{q} g(\delta,q),$$

using Lemma 3.2 with $X_k := q_k$ and $y_k := \sum_j p_j P_{kj}^{\frac{1}{1+\delta}}$, a final expression for the reliability rate function E(R) is:

$$E(R) = \max_{p \in \mathbb{P}^n} \max_{\delta \geq 0} \left\{-\delta R - \log \sum_k \left\{\sum_j p_j P_{kj}^{\frac{1}{1+\delta}}\right\}^{1+\delta}\right\} \qquad (21)$$

This result coincides with Theorem 18 given in (Ref. 14). The second term in (21) is the so-called Gallager function. The dual representation (21) is useful for deriving efficient computational algorithms, see e.g., (Ref. 20).

REFERENCES

[1] Kullback, S., Leibler, R.A., "On Information and Sufficiency", Ann. Math. Stat., Vol. 22, pp. 79-86, 1951.

[2] Kullback, S., Information Theory and Statistics, John Wiley & Sons, Inc., New York, 1959.

[3] Akaike, H., "Information Theory and an Extension of the Maximum Likelihood Principle", Proc. of the 2nd Int. Symp. on Information Theory, 1972.

[4] Charnes, A., Raike, W.M., and Bettinger, C.O., "An Extremal and Information Theoretic Characterization of some Interzonal Transfers," Socio-Economic Planning Sciences, Vol. 6, pp. 531-537, 1972.

[5] Charnes, A., Cooper, W.W., and Learner, D.B., "Constrained Information Theoretic Characterizations in Consumer Purchase Behavior", Journal of the Operational Research Society, pp. 833-840, 1978.

[6] Charnes, A., Cooper, W.W., "An Extremal Principle for Accounting Balance of a Resource Value Transfer Economy: Existence, Uniqueness and Computations", Acad. Naz. Lincei, Series VIII, Vol. LVI, pp. 556-561, 1974.

[7] Charnes, A. and Cooper, W.W., "Constrained Kullback-Leibler Estimation; Generalized Cobb-Douglas Balance, and Unconstrained Convex Programming", Acad. Naz. Lincei, Series VIII, Vol. LVIII, pp. 568-576, 1975.

[8] Shannon, C.E., "A Mathematical Theory of Communication," Bell Syst. Tech. Journal, Vol. 27, pp. 379-423, pp. 623-656, 1948.

[9] Ben-Tal, A., Charnes, A., "A Dual Optimization Framework for some Problems of Information Theory and Statistics", Problems of Control and Information Theory, Vol. 8, pp. 387-401, 1979.

[10] Gallager, R.G., Information Theory and Reliable Communication, J. Wiley and Sons, Inc., New York, 1968.

[11] Jelinek, F., Probabilistic Information Theory, McGraw-Hill, New York, 1968.

[12] Gallager, R.G., "A Simple Derivation of the Coding Theorem and Some Applications", IEEE Trans. Information Theory, Vol. IT-11, pp. 3-18, 1965.

[13] Csiszar, I. and Korner, J., Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic, New York, 1981.

REFFRENCES (cont'd)

[14] Blahut, R.E., "On Hypothesis Testing and Information Theory", IEEE Trans. Information Theory, Vol. IT-20, pp. 405-417, 1974.

[15] Rockafellar, R.T., 'Duality and Stability in Extremum Problems Involving Convex Functions", Pacific J. Math., 21, 1, pp. 167-186, 1976.

[16] Rockafellar, R.T., "Conjugate Duality and Optimization", Regional Conference Series in Applied Mathematics, SIAM No. 16, 1974.

[17] Laurent, P.J., Optimisation et Approximation", Hermann, Paris, 1972.

[18] Ponstein, J., Approaches to the Theory of Optimization, Cambridge University Press, 1980.

[19] Rockafellar, R.T., "Minimax Theorems and Conjugate Saddle Functions", Math. Scand., Vol. 14, pp. 151-173, 1964.

[20] Arimoto, S., "Computation of Random Coding Exponent Functions", IEEE Trans. Information Theory, Vol. IT-22, pp. 665-671, 1976.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER CCS 497	2. GOVT ACCESSION NO. A152 336	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) The Role of Duality in Optimization Problems Involving Entropy Functionals, with Applications to Information Theory		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A. Ben-Tal, M. Teboulle, A. Charnes		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0295 N00014-81-C-0236
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Cybernetic Studies The University of Texas at Austin Austin, Texas 78712		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research (Code 434) Washington, D.C.		12. REPORT DATE September 1984 13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)		15. SECURITY CLASS. (*of this report*) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

Optimization in Infinite Dimensional Spaces; Duality in Convex Optimization; Entropy; divergence; Information Theory; Channel Capacity; Reliability Rate Function; Error Exponent Function.

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

We consider infinite dimensional optimization problems involving entropy-type functionals in the objective function as well as in the constraints. A duality theory is developed for such problems and applied to the reliability rate function problem in Information Theory.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)